Volume 10 • Number 4 • 2007
VALUE IN HEALTH

Validation of a Condition-Specific Measure for Women Having an Abnormal Screening Mammography

John Brodersen, MD, GP, PhD, Hanne Thorsen, MD, PhD, Svend Kreiner, MSc

Institute of Public Health, University of Copenhagen, Copenhagen, Denmark

ABSTRACT

Objectives: The aim of this study is to assess the validity of a new condition-specific instrument measuring psychosocial consequences of abnormal screening mammography (PCQ-DK33).

Methods: The draft version of the PCQ-DK33 was completed on two occasions by 184 women who had received an abnormal screening mammography and on one occasion by 240 women who had received a normal screening result. Item Response Theories and Classical Test Theories were used to analyze data. Construct validity, concurrent validity, known group validity, objectivity and reliability were established by item analysis examining the fit between item responses and Rasch models.

Results: Six dimensions covering anxiety, behavioral impact, sense of dejection, impact on sleep, breast examination, and sexuality were identified. One item belonging to the dejection dimension had uniform differential item functioning. Two items not fitting the Rasch models were retained because of high face validity. A sick leave item added useful information when measuring side effects and socioeconomic consequences of breast cancer screening. Five "poor items" were identified and should be deleted from the final instrument.

Conclusions: Preliminary evidence for a valid and reliable condition-specific measure for women having an abnormal screening mammography was established. The measure includes 27 "good" items measuring different attributes of the same overall latent structure—the psychosocial consequences of abnormal screening mammography.

Keywords: adverse effects, false-positive reactions, mass screening, questionnaire design.

Introduction

If a woman follows the European Union-recommended biannual breast cancer screening program for 20 years her lifetime risk for a false-positive screening mammography will be 20% to 25%, possibly even higher [1]. In the UK, where women aged 50–65 years are offered screening every 3 years, more than 50,000 women per year will receive a false-positive screening mammography (hereafter referred to as a false positive) [2]. False positives cause significant adverse consequences including costly follow-up tests and an increased use of health-care services [3–5]. In addition, numerous studies have shown that women recalled for further investigations after an abnormal screening mammography, later confirmed as a false positive, experience significant adverse psychosocial effects [6]. In these studies a variety of questionnaires have been used to measure the adverse effects. Except for one questionnaire most of the instruments have been developed for other purposes and have a more or less generic character [7].

Address correspondence to: John Brodersen, Department of General Practice, Institute of Public Health, University of Copenhagen, Øster Farimagsgade 5, 24Q, Postbox 2099, DK-1014 Copenhagen, Denmark. E-mail: j.brodersen@gpmed.ku.dk

10.1111/j.1524-4733.2007.00184.x

A requirement for the validity of a questionnaire is that it has high content relevance and high content coverage [8]. A condition-specific measure insures higher content coverage compared with measures developed for generic conditions [9]. A review has shown that some of the most frequently used generic measures in the setting of breast cancer screening: the General Health Questionnaire, the Hospital Anxiety and Depression Scale and the State-Trait Anxiety Inventory, have problems with language, content relevance and content coverage [7]. McCaffery and Barrat underline the importance of high content validity when using questionnaires to measure psychosocial consequences of screening [10].

The Psychological Consequences Questionnaire (PCQ) was developed in 1992 by Jill Cockburn to measure the short-term psychosocial consequences of the actual act of participation in breast cancer screening. It consists of 12 items covering negative aspects and 10 items covering positive aspects of participation [11]. The full PCQ (negative and positive items) has been used in only one study of the process of participating in breast cancer screening [12]. The negative items have predominantly been used in studies of the adverse consequences of false positive and not of the consequences of the actual participation [6]. Nevertheless, the content validity of the PCQ has never been

tested in the setting of abnormal and false-positive screening mammography [7].

When summing raw scores of items in a scale an assumption of unidimensionality is made, that is, the items describe different aspects of the same construct and can be added [13,14]. When the response options are categorical on an ordinal scale as in many questionnaires Item Response Theory (IRT) and Rasch models can be used to assess the psychometric properties of the questionnaire. The Rasch models provide formal representation of perfect measurement. Where items are shown to fit a Rasch model the measure can be shown to posses criterion-related construct validity [15], to be objective [16], sufficient [17], and therefore also reliable [18]. Measurements are specific objective if comparisons of measurements do not systematically depend on arbitrary choices made in connection with measurement. The choice of items is, for example, best regarded as an arbitrary choice of item from a larger item bank. Specific objectivity thus requires that comparison of two persons by one version of a scale does not differ systematically from comparisons using another similar set of valid items. Similarly, comparison of two persons should not depend on measurements on other persons [16]. For these reasons the Rasch model is considered a valuable "gold standard" against which measures should be compared. Reliability and different aspects of validity can also be assessed using Classical Test Theory (CTT) [8]. The relation between CTT and IRT has been described by Holland and Hoskens [19] and it may be an advantage to combine the two theories. Item analyses by Rasch models explore in-depth the degree to which the requirements of construct validity are met. Items are assumed to monotonically relate to one dimension and they are assumed to be locally independent. It is also assumed that there is no differential item functioning (DIF), that is, where an item functions differently in subpopulations such as in an intervention group and a control group [15,20]. The sufficiency of the model support computation of the raw scores. IRT analyses also explore how the items included in each dimension are interrelated and ordered on a latent trait (e.g., psychosocial consequences of a false positive) [21].

The strength of analyses based on the Rasch Model is that the model is build on preassumptions closer to reality than analyses based on CTT [14,21]. The Rasch models describe how item responses depend both on person and item parameters. The person parameter is assumed to be unidimensional but item parameters may be multidimensional when item responses are ordinal categories.

The purpose of this study was to validate a new condition-specific instrument measuring psychosocial consequences of abnormal screening mammography (PCQ-DK33) using both IRT and CTT.

Materials and Methods

A qualitative study to assess the content validity of the PCQ was conducted in a setting of abnormal screening mammography. The qualitative study highlighted the need to make radical changes to the questionnaire if it was to be used in this setting. Therefore, the draft version of the questionnaire statistically tested in the present study can be regarded as a new condition-specific instrument with 33 items (PCQ-DK33, see the summary in Table 4). Data were collected at two screening centers: the Copenhagen University Hospital and Odense University Hospital.

Group 1–Time 1

Over a period of 20 weeks from November 2002 all women who were recalled because of an abnormal screening mammography were consecutively included in the study. Before any further examinations to establish if the abnormal screening result was true or false the women were asked to complete two questionnaires: 1) the PCQ-DK33 with the items randomly ordered, and 2) the Danish version of the Nottingham Health Profile (NHP).

The NHP is a questionnaire measuring health status. It was originally developed in the UK [22] and has been adapted into a large number of languages including Danish [23,24]. The NHP consists of six sections covering energy, pain, emotional reaction, sleep, social isolation, and physical mobility. It was selected as a comparator to assess concurrent validity [8] for the present study because emotional reactions and sleep problems are well covered by the measure. It was hypothesized that the new instrument measuring psychosocial consequences of abnormal screening mammography would converge with the emotional section of the NHP and diverges from the pain and mobility sections.

Group 1–Time II

Two weeks after having completed the first questionnaire package the women were sent the draft version of the PCQ-DK33. At this point most of the women would know whether their abnormal screening result was false positive or whether they had breast cancer. They were asked to complete the questionnaire and return it in an enclosed stamped addressed envelope. It was hypothesized that women received a diagnosis of breast cancer experienced more severe psychosocial consequences than women with a known false-positive screening result (known group validity and responsiveness) [8,25]. It was also hypothesized that there would be a decrease in the negative psychosocial consequences from abnormal to known false-positive screening result.

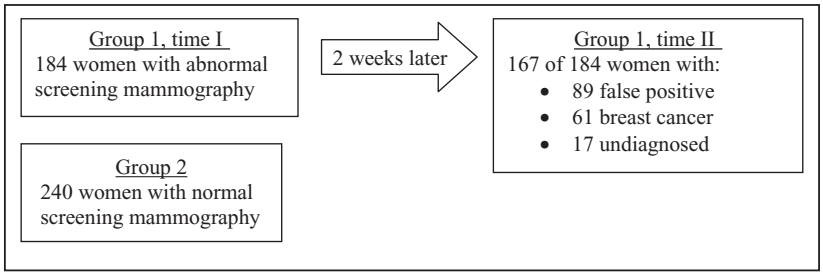

Figure 1 Data collection of the PCQ-DK33.

Group 2

For each woman in Group 1 who had completed the questionnaires at time I another two women were sent the draft version of the PCQ-DK33 and asked to return the completed questionnaire in an enclosed stamped addressed envelope. These women had had a normal screening mammography at the same time and at the same clinic as the women in Group 1–Time I. The recruitment procedure for women in Group 2 was: if the screening mammography of a woman was abnormal then the two women having a normal screening mammography and being screened just before and after this woman were included in Group 2. In contrast to Group 1, where informed consent was obtained at the recall clinics, it was not possible to obtained informed consent before posting the test questionnaire to the women in Group 2. For ethical and legal reasons the test questionnaire was posted to the women by the screening clinics anonymously and only the age of the participants was disclosed for the researchers. Therefore, it was not possible to send reminders to the women in Group 2. It was hypothesized that women with an abnormal screening result experienced more severe psychosocial consequences than women with a normal screening result (known group validity) [8,25].

In Group 1–Time I, 220 women were eligible. Of these, 16 (7.3%) were not invited to participate because of sick leave among the clinic staff. Of those asked to participate 90.2% agreed to complete the questionnaires. At Time II, 90.8% returned the PCQ-DK33 after one reminder. In Group 2, 400 women received the questionnaire by post and 60% were returned. There were no statistically significant differences in mean and range of age between the women in the three groups. Figure 1 illustrates the data collection including the numbers of women in each subgroup.

Three questionnaires were returned without being completed. Among the remaining 588 questionnaires 0.3% to 1.9% randomly distributed missing values per item were observed.

Item responses were analyzed by the conditional distribution of items given total person scores to avoid assumptions on the distribution of the latent trait being measured. The pair-wise estimation procedure implemented in software program RUMM2020 was used to estimate the item parameters [26,27]. The analysis of the fit of item responses to the Rasch model were based on analyses of residuals comparing observed to expected item responses, both for separate individuals and for different score groups. The overall fit of the model was assessed by the Wright-Panchapakesan chi-square statistic summarizing standardized residuals over score groups and items [28]. Item fit statistics summarizing standardized residuals in different score groups were used to identify misfitting items. Data from Group 1–Time I and Time II and data from Group 2 were pooled for IRT analyses. Nevertheless, the sick leave item was not included in the Rasch analyses because the response options differed entirely from the response options in the remaining 32 items. DIF relative to person covariates was checked by analyses of variance examining the degree to which individual residuals for specific items depended on the covariates. Absence of evidence of interaction between the covariates and the estimated trait parameters were taken as evidence of DIF being uniform. A small subgroup of 17 women undiagnosed 2 weeks after their abnormal screening mammography was not included in the analyses of DIF.

Finally, the assumption of local independence was checked by examination of the degree to which individual residuals were correlated.

Reliability was assessed by Cronbach's alpha defining the lower bound for the test–retest correlation of the raw scores [29,30] and by the so-called Person Separation Index calculating the lower bound for the test–retest correlation of the estimated values of the latent trait being measured [31].

Data were analyzed with CTT by using the software SPSS for Windows version 13.0 (SPSS Inc., Chicago, IL, USA) and the software Mplus 2.14 for confirmatory factor analysis [32]. For IRT analyses the software RUMM2020 was used [26].

The study was approved by the local ethical committee.

Results

The initial item analysis of the complete set of 32 items rejected the Rasch model. Strong evidence of local dependence indicated that the PCQ-DK33 was not unidimensional. The subsequent separate analyses

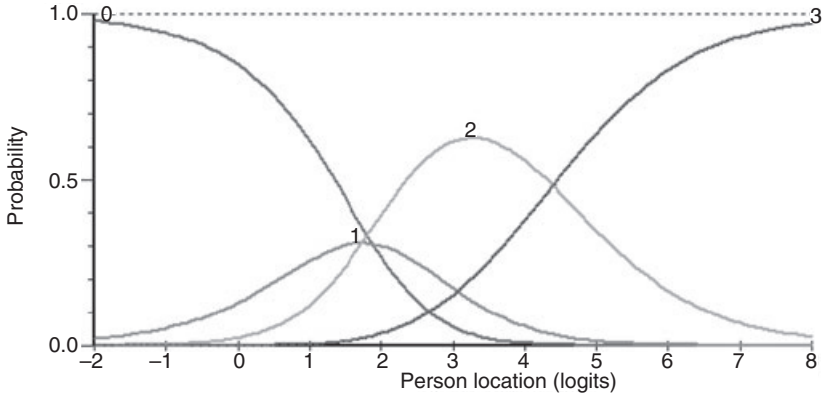

Figure 2 Category probability curve of item 29 "terrified." The logit scale from –2 to +8 on the *x*-axis symbolizes the latent trait of anxiety, with the severity of anxiety increasing toward the right. The *y*-axis symbolizes the probability of affirming the response categories: 0 "not at all," 1 "a bit," 2 "quite a bit," and 3 "a lot."

confirmed the multidimensionality expected from the qualitative study on face and content validity of the instrument.

Six items covering anxiety formed one dimension and none of these items had DIF in either of the subgroups. In one of the six items, "felt terrified" (No. 29), the thresholds of the response categories were not in order (Fig. 2). Of eight items covering the impact on behavior after abnormal screening mammography, seven items fitted the Rasch model and no DIF was observed. Among the seven behavioral items the thresholds of item "difficulty doing everyday things around the house" (No. 28) were not in order (Fig. 3). Six items describing the sense of dejection and sadness after abnormal screening mammography fitted the Rasch model forming one dimension. Nevertheless, the item "felt sad" (No. 14) had uniform DIF in two of four subgroups as shown in Figure 4. After deleting this item the five remaining items still fitted the Rasch model with all thresholds in order and no DIF.

From a content perspective the Rasch analyses confirmed three more dimensions each with two items. These three dimensions described impact on breast examination, sleep and sexuality. Nevertheless, in the breast examination dimension both items had uniform DIF. The item "examined my breasts" (No. 16) had uniform DIF in the group received a diagnosis of breast cancer compared with the remaining groups (Fig. 5) and the item "examined my breasts in the mirror" (No. 21) had uniform DIF in the group with normal screening mammography compared with the remaining groups (Fig. 6). In the sexuality dimension the thresholds of the item "less interest in sex" (No. 31) were not in order (Fig. 7).

Besides testing the six identified dimensions for DIF in the sampled groups (Fig. 1), DIF was also tested for age and screening center. These tests showed no DIF.

There was no local dependency among the items in the identified dimensions.

The two sleeping items fitting the Rasch model and the item "taking sleeping tablets" are from a content point of view equal to three of the five items included in the sleep section of the NHP. Rasch analyses of the sleep section of the NHP showed that four of the five sleeping items fitted the Rasch model except for the sleeping tablet item.

The Wright-Panchapakesan chi-square fit statistics, the Person Separation Index and the Cronbach's alpha of the six dimensions fitting the Rasch model in the PCQ-DK33 are listed in Table 1. The fit statistic of the behavioral subscale is marginally significant ($P = 0.039$). Adjusting *P*-values to control the false

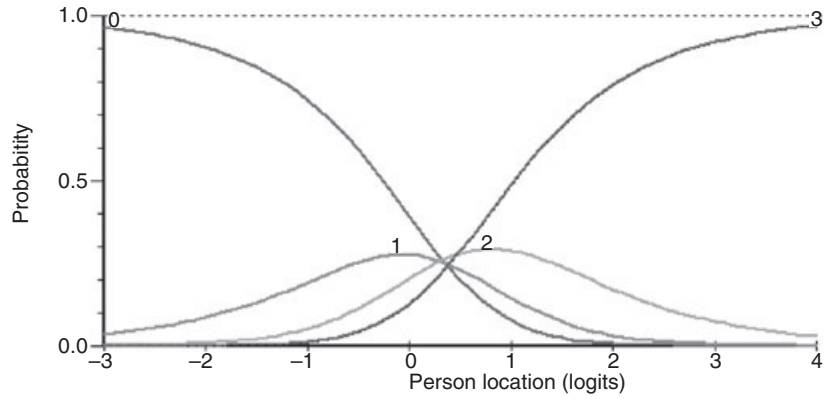

Figure 3 Category probability curve of item 28 "difficulty doing things around the house." The logit scale from –3 to +4 on the *x*-axis symbolizes the latent trait of behavioral impact, with the severity of negative impact increasing toward the right. The *y*-axis symbolizes the probability of affirming the response categories: 0 "not at all," 1 "a bit," 2 "quite a bit," and 3 "a lot."

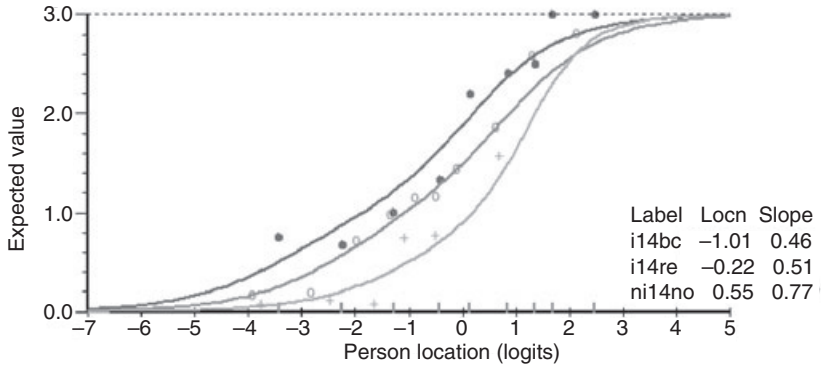

Figure 4 Item characteristics curves (ICC) of item 14 "felt sad" showing uniform differential item functioning (no statistical significant difference between the slopes of the three ICC) between the group of women received a diagnosis of breast cancer (i14bc), the group of women having normal screening mammography (ni14no) and the remaining women – women having abnormal and a false-positive screening mammography (i14re). The logit scale from –7 to +5 on the *x*-axis symbolizes the latent trait of sense of dejection, with the severity of dejection increasing toward the right. The *y*-axis symbolizes the values of responses options: 0 "not at all," 1 "a bit," 2 "quite a bit," and 3 "a lot." Label symbolizes the ICC for the three subgroups and Locn symbolizes the item location.

discovery rate and so avoid spurious significant results due to multiple testing suggested that the result should be regarded as insignificant [33].

For each subgroup Table 2 shows the mean score, the standard error of mean and the standard deviation for all six dimensions fitting the Rasch model.

Items numbered 1, 7, 9, 10, 25, 27, and 30 did not fit the Rasch model. The face validity of these items was checked by re-auditing tape recordings from the focus group interviews conducted during the adaptation of the PCQ into Danish. The item "less attractive" (No. 1) and the item "busy to take mind off things" (No. 10) had significant face validity for women who had had surgery or had been on early recall after abnormal screening mammography.

Cronbach's alpha increased when deleting items 7, 25, 27, and 30 and dropped when deleting items 1, 9, and 10. This indicated that items 7, 25, 27, and 30 were "poor" items because alpha is expected to decrease when valid items are deleted from a summated score.

A subsequent confirmatory factor analysis was conducted to estimate how the six dimensions and the three single items (No's 1, 10, and 14) were correlated. The analysis revealed a positive correlation between all six dimensions and the three single items. Only very weak evidence was disclosed against a model assuming that one latent trait lies behind the six dimensions and the three single items of the PCQ-DK33 ($P = 0.0463$).

When testing for known group validity a statistically significant difference was found between women having an abnormal screening mammography and those having a normal screening mammography. In all six dimensions and in the three single items the *P*-value of the Pearson chi-square was less than 0.0005. In the single item about sick leave (No. 33) the *P*-value was 0.026.

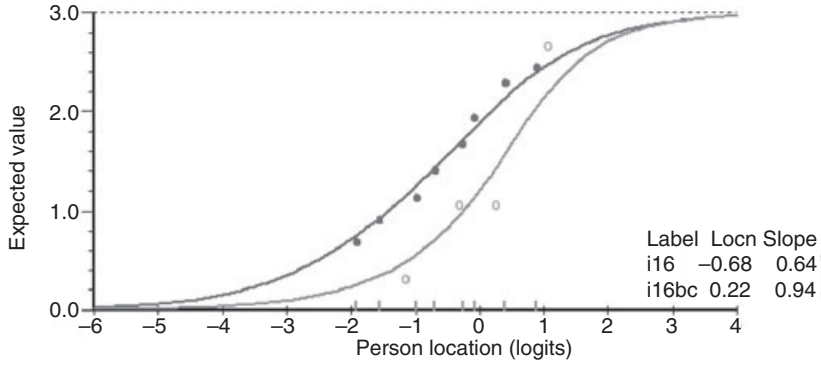

Figure 5 Item characteristics curves (ICC) of item 16 "examined my breasts" showing uniform differential item function (no statistical significant difference between the slopes of the two ICC) between the group of women received a diagnosis of breast cancer (i16bc) and the remaining women – women having an abnormal, a false positive and a normal screening mammography (i16re). The logit scale from –6 to +4 on the *x*-axis symbolizes the latent trait of breast examination, with the severity of breast examination increasing toward the right. The *y*-axis symbolizes the values of responses options: 0 "not at all," 1 "a bit," 2 "quite a bit," and 3 "a lot." Label symbolizes the ICC for the three subgroups and Locn symbolizes the item location.

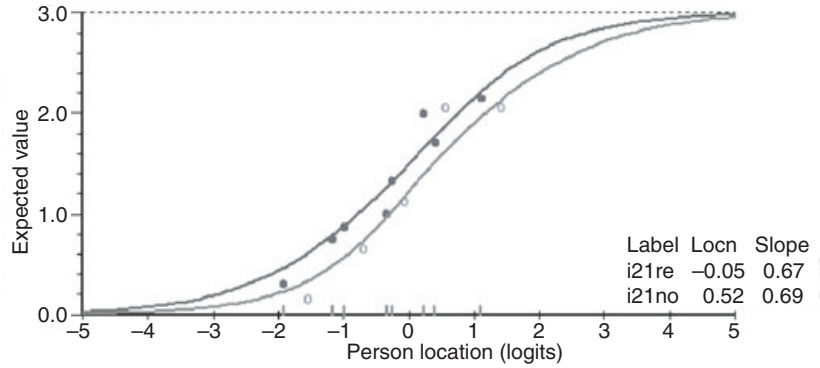

Figure 6 Item characteristics curves (ICC) of item 21 "examined my breasts in the mirror" showing uniform differential item functioning (no statistical significant difference between the slopes of the three ICC) between the group of women with a normal screening mammography (i21no) and the remaining women—women having an abnormal, a false positive and true positive (breast cancer) screening mammography (i21re). The logit scale from –5 to +5 on the *x*-axis symbolizes the latent trait of breast examination, with the severity of breast examination increasing toward the right. The *y*-axis symbolizes the values of responses options: 0 "not at all," 1 "a bit," 2 "quite a bit," and 3 "a lot." Label symbolizes the ICC for the three subgroups and Locn symbolizes the item location.

There was a statistically significant difference between women received a diagnosis of breast cancer and those having a known false positive in five of the six dimensions and in the three single items (No's 1, 10, and 33) with the highest *P*-value as 0.016. The sexuality subscale could not differ between women received a diagnosis of breast cancer and those having a known false-positive screening result. Additional analyses on the sexuality subscales found no statistically significant difference between women having an abnormal screening mammography and those received a diagnosis of breast cancer. They also showed no difference between women having an abnormal screening mammography and those with a known false positive.

As a test of concurrent validity the Pearson correlation was established between the sections of the NHP and all the 27 "good" items (24 items in the six identified dimensions and items 1, 10, and 14). The Pearson correlation was also established for each of

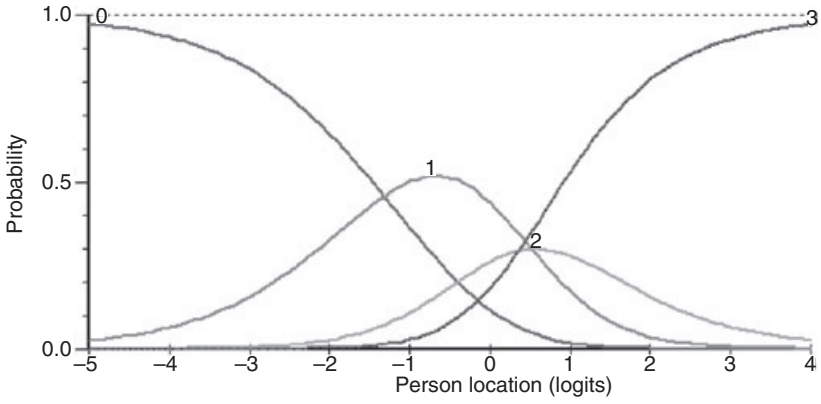

Figure 7 Category probability curve of item 32 "not felt like having my breast caressed. The logit scale from –5 to +4 on the *x*-axis symbolizes the latent trait of sexuality, with the severity of negative impact of sexuality increasing toward the right. The *y*-axis symbolizes the probability of affirming the response categories: 0 "not at all," 1 "a bit," 2 "quite a bit," and 3 "a lot."

Table 1 Wright-Panchapakesan (WP) fit statistics, Person Separation Index and the Cronbach's alpha of six dimensions in the PCQ-DK33

Dimensions (number of items)	WP χ^2	Degrees of freedom	*P*-value	Person separation index	Cronbach's alpha
Anxiety (6)	61.81	52	0.166	0.94	0.92
Behavioral (7)	67.71	49	0.039	0.88	0.86
Sense of dejection (5)	55.14	40	0.056	0.93	0.89
Sleep (2)	8.49	9	0.486	0.89	0.90
Breast examination (2 or 4*)	19.60 (28.34*)	8 (21*)	0.01 (0.131*)	0.68 (0.70*)	0.71
Sexuality (2)	7.06	10	0.720	0.81	0.83

*After item split according to the uniform differential item functioning found.

Table 2 Mean scores, standard error of mean and standard deviation of all six dimensions

	Anxiety					Behavioral					Dejection				
Dimensions	A	B	C	D	E	A	B	C	D	E	A	B	C	D	E
No*	178	52	97	16	232	173	50	97	16	231	179	52	96	15	234
Mean	6.39	7.38	3.04	4.38	0.97	3.53	4.96	1.72	3.63	0.47	5.38	6.10	2.25	3.73	1.03
SE mean	0.36	0.73	0.42	1.24	0.13	0.28	0.61	0.27	1.04	0.11	0.29	0.60	0.34	1.14	0.12
SD	4.77	5.30	4.18	4.95	2.00	3.68	5.30	2.63	4.16	1.63	3.94	4.33	3.29	4.42	1.84
	Sleep					Breast examination					Sexuality				
Dimensions	A	B	C	D	E	A	B	C	D	E	A	B	C	D	E
No*	179	52	97	16	235	181	52	97	16	235	173	50	97	16	235
Mean	2.10	2.48	1.10	1.63	0.33	2.2†	1.7†	1.1†	1.1	0.7†	0.84	1.08	0.62	0.94	0.15
SE mean	0.15	0.34	0.18	0.46	0.06	0.1‡	0.2‡	0.1‡	0.36	0.1‡	0.11	0.26	0.15	0.46	0.04
SD	2.00	2.43	1.73	1.82	0.94				1.44		1.50	1.82	1.43	1.84	0.67

*Number of subjects included in the Rasch analyses
†Mean scores are estimated from the means of the person locations on the latent trait according to the uniform differential item functioning found in items 16 and 21.
‡The SE mean are estimated from the SE means of the person locations.
A, Group 1–Time I; B, Group 1–Time II, women received a diagnosis of breast cancer; C, Group 1–Time II, women with known false positive; D, Group 1–Time II, women undiagnosed; E, Group 2.

the six Rasch-fitting dimensions. The results confirmed the hypothesis made before the analysis (Table 3).

A summary of the results from the psychometric analyses of the PCQ-DK33 are given in Table 4.

Discussion

Six Rasch-fitting dimensions were identified encompassing 24 items. The dimensions cover: anxiety, behavioral impact, sense of dejection, impact on sleep, breast examination, and sexuality. The six dimensions had sufficient sensitivity to distinguish between groups a priori hypothesized to be different. The correlations between the six Rasch-fitting dimensions and the six sections of the NHP were also as could be expected from a content point of view.

Two items not fitting the Rasch models were retained in the questionnaire because of high face validity. These were concerned with "feeling less attractive" and "kept busy to take mind off things." An item "felt sad" belonged to the dejection dimension but had uniform DIF relative to diagnostic subgroups. The six dimensions and these three single items converged as expected with the emotional section of the NHP and diverged as expected with the pain and the physical mobility sections.

The sick leave item was not included in the original Australian version of the PCQ. Nevertheless, it seems to add useful information when measuring side effects and socioeconomic consequences of breast cancer screening.

Five "poor items" were identified. Four of these misfitted the Rasch model and had low face validity. Cronbach's alpha also increased when they were deleted. The fifth "poor" item "being tired" showed an unclear picture. Although, it did not fit the Rasch model and had low face validity, Cronbach's alpha dropped when it was deleted. Perhaps "being tired" describes a too general condition. Therefore, it is suggested that the five "poor" items should be deleted from the final instrument. It is worth mentioning that one of the "poor" items "keeping things from those who are close to you" is an item belonging to the original Australian version of the PCQ [11].

Collecting questionnaire data under two different conditions may result in biases. At time I, the women completed the PCQ-DK33 at the recall clinic. Two weeks later at time II the same women completed the PCQ-DK33 at home. Some women had the additional examinations at the recall clinic only one day after receiving the letter about the abnormal screening mammography. This short time interval made it nec-

Table 3 Concurrent validity (convergent and divergent validity) of 27 "good" items and the six identified dimensions of the PCQ-DK33 and the sections of the NHP at time I–group 1

NHP sections	27 "good" items	Anxiety	Behavioral	Dejection	Sleep	Breast examination	Sexuality
Energy level	0.43	0.30	0.49	0.37	0.30	–0.05	0.30
Pain	0.15	0.15	0.14	0.13	0.14	–0.13	0.09
Emotional reaction	0.70	0.65	0.68	0.67	0.55	0.15	0.35
Sleep	0.48	0.38	0.40	0.39	0.69	0.12	0.20
Social isolation	0.26	0.27	0.29	0.24	0.25	0.01	0.11
Physical mobility	0.16	0.10	0.17	0.10	0.10	–0.17	0.13

The correlations are calculated as coefficients from the Pearson correlation.
NHP, Nottingham Health Profile.

Table 4 Summary of result from the psychometric analyses of the PCQ-DK33

The items of the PCQ-DK33 in order of appearance in the draft version	Misfit to the Rasch model or subscale belonging	Probability of fit to the Rasch model	Degrees of freedom	χ^2	Cronbach's alpha	Face validity	Single or "poor" item	Final decision
1. Less attractive	Misfit	0.093	7	12.241	Dropped	High	Single item	Inclusion
2. Worried	Dejection	0.471	8	7.627				Inclusion
3. Worried about my future	Anxiety	0.280	9	10.931				Inclusion
4. Scared	Anxiety	0.208	9	12.098				Inclusion
5. Irritable	Behavioral	0.513	7	6.236				Inclusion
6. Quieter than normal	Behavioral	0.902	7	2.805				Inclusion
7. Keeping things from those who are close to you	Misfit	<0.0005	8	50.659	Increased	Low	"Poor" item	Exclusion
8. Slept badly	Sleep	0.524	4	3.209				Inclusion
9. Tired	Misfit	0.003	5	21.708	Dropped	Low	"Poor" item	Exclusion
10. Busy to take mind off things	Misfit	<0.0005	8	54.375	Dropped	High	Single item	Inclusion
11. Hard to concentrate	Behavioral	0.189	7	9.989				Inclusion
12. Time passed slowly	Dejection	0.852	8	4.051				Inclusion
13. Change in appetite	Behavioral	0.426	7	7.024				Inclusion
14. Sad	Dejection*	0.016	8	18.855				Inclusion
15. Upset	Anxiety	0.282	9	10.907				Inclusion
16. Examined my breasts	Breast examination	0.152	4	6.711				Inclusion
17. Restless	Anxiety	0.262	9	10.039			Single item?	Inclusion
18. Nervous	Anxiety	0.244	9	11.486				Inclusion
19. Uneasy	Dejection	0.002†	8	24.413				Inclusion
20. Taken long time to fall asleep	Sleep	0.383	5	5.277				Inclusion
21. Examined my breasts in the mirror	Breast examination	0.012	4	12.891				Inclusion
22. Withdrawn into myself	Behavioral	0.060	7	13.532				Inclusion
23. Unable to cope	Dejection	0.388	8	8.483				Inclusion
24. Depressed	Dejection	0.228	8	10.564				Inclusion
25. Symptoms from the breast (pins and needles)	Misfit	<0.0005	5	26.942	Increased	Low	"Poor" item	Exclusion
26. Difficulty dealing work or other commitments	Behavioral	0.006†	7	19.862				Inclusion
27. Headache	Misfit	0.004	6	19.379	Increased	Low	"Poor" item	Exclusion
28. Difficulty doing things around the house	Behavioral	0.310	7	8.261				Inclusion
29. Terrified	Anxiety	0.609	9	6.344				Inclusion
30. Taking sleeping tablets	Misfit	0.003	5	17.688	Increased	Low	"Poor" item	Exclusion
31. Less interest in sex	Sexuality	0.828	5	2.154				Inclusion
32. Not felt like having my breast caressed	Sexuality	0.428	5	4.905				Inclusion
33. Sick leave	Not included	Not included	Not included	Not included	Not included	High	Single item	Inclusion

*Item was deleted from the dejection dimension because of differential item function.
†Marginal misfit also after a correction of Benjamini-Hochberg procedure [33].
The item (No. 33) asking about sick leave was not included in the Rasch analyses.

essary to complete the questionnaires at the recall clinic instead of receiving it by post. Women completing a questionnaire at a clinic may "smarten up" their answers to be polite. Therefore, the negative psychosocial consequences of an abnormal screening mammography would most likely be underestimated.

Three items showed DIF relative to diagnostic subgroups. If an item functions differently in subpopulations other psychometric properties should determine the "destiny" of the item. The qualitative study preceding the present study showed that the two items forming the breast examination dimension cover an important area in the context of abnormal and false-positive screening mammography. The results from the concurrent validity tests confirmed that the breast examination dimension measured something different from the other five dimensions. Consequently, it would be unwise to remove these items. Nevertheless, special precautions should be taken when calculating scores of this dimension. The content of the item "being sad" is close to the content of the other five items in the dejection dimension. Therefore, it seems reasonable to delete this item if future studies continuously show DIF.

Three items from three different dimensions had problems with the order of thresholds. As seen in Figures 2, 3 and 7 the disorder was caused by minor problems. If future studies show the same pattern it has to be decided whether the response categories "A bit" and "Quite a bit" should be merged either by rescoring the items or redesigning the layout.

Two items belonging to two different dimensions showed a marginal misfit of probability to the Rasch model (No.'s 19 and 26, Table 4). Nevertheless, the overall fit of the dimensions were satisfactory. Future studies including the questionnaire will show whether these findings are consistent.

It was surprising that the sexuality dimension only distinguished between women having a normal and women having an abnormal screening mammography. This may indicate that the negative impact on sexuality after having an abnormal screening result had not declined or vanished 1 or 2 weeks after women were "free from" cancer suspicion.

The convergence between the dimensions of anxiety, behavior and dejection and the emotional section of the NHP indicates some overlap between these three dimensions. Only longitudinal studies will reveal whether this overlap is caused by redundancy. Nevertheless, removing any of the dimensions would decrease content coverage. The lack of convergence between the two dimensions "breast examination" and "sexuality" and the emotional section of the NHP contradicts redundancy among the six dimensions.

The establishment of a traditional test–retest reliability coefficient requires at least 2 to 4 weeks where the condition for the respondents is stable [8,25]. The condition for the women in group 1 changed dramatically from time I to time II. At time I all women had been told that their screening result was abnormal. At time II nearly all women knew their diagnosis: breast cancer or false positive. A satisfactory reliability of the measure was assessed with Cronbach's alpha and Person Separation Index (Table 1).

As shown in the Rasch analyses; four of five NHP sleep-items fitted the model. Two of the four items are content-wise equivalent to the two Rasch-fitting items of the PCQ-DK33. Therefore, adapting the two non-equivalent sleep items from the NHP would probably add nuances to the sleep dimension of the new instrument.

As in many other questionnaires, the response options of the PCQ-DK33 are ordered categories. Several models for ordinal categorical responses have been suggested. The model used in the present study is the partial credit model (PCM) in which the item parameters are sometimes described as threshold parameters [34]. The thresholds in the PCM may differ between items. In contrast, the rating scale model assumes that thresholds are homogenous across items apart from an additive factor describing the item difficulty [35]. Rating scale models were considered during the analysis but abandoned because of lack of fit between the model and the observed item responses.

The present study has shown the advantages of combining analyses based on IRT and CTT when assessing the psychometric properties of a questionnaire including dimensionality and "good" and "poor" single items. After establishing unidimensionality with the Rasch model CTT analyses were subsequently conducted. This order of analyses had several advantages: First, more than half of the Rasch-misfitting items were confirmed also to be "poor" by the analyses of Cronbach's alpha. Second, the internal consistency expressed as a Cronbach's alpha coefficient was calculated only on items included in the Rasch-fitting dimensions. Third, the results of testing known group validity and concurrent validity were only established on "good" items.

Conclusion and Perspectives

In conclusion, the reliability and the construct validity of a condition-specific measure with high content validity for women having an abnormal screening mammography have been preliminary demonstrated. This new questionnaire covers the impact of experiencing an abnormal screening mammography on: anxiety, behavior, dejection, sexuality, sleep, and breast examination. In addition, the measure includes three single items covering: sick leave, feeling less attractive, and kept busy to take mind off things.

The new instrument is currently in use in a major Danish survey and has been translated into Dutch,

English, and Norwegian. Future analyses on data from surveys will be hoped to give an answer to the questions left from the present study:

- Should all the three single items be retained in the final version of the measure?
- Do the four items covering sleep form one dimension?
- Will the item "felt sad" still have uniform DIF?
- Will some items still have thresholds that are not in order?

We are in great debt to the late professor Jill Cockburn for her inspiration and strong supported of developing a condition-specific measure of psychosocial consequences of abnormal and false-positive screening mammography. We want to thank the staff at the breast cancer screening clinic and the recall clinic at the Copenhagen University Hospital and at the Odense University Hospital for recruiting women to this study.

Source of financial support: The project has been sponsored by The Psychosocial Cancer Foundation; Denmark, The National Health Insurance; Denmark, Praktiserende Laegers Uddannnelses og Udviklingsfond (The General Practitioners Education and Development Foundation); Denmark. There are no known conflicts of interest for the authors.

References

1 Elmore JG, Barton MB, Moceri VM, et al. Ten-year risk of false positive screening mammograms and clinical breast examinations [see comments]. N Engl J Med 1998;338:1089–96.
2 NHS. NHS Cancer Screening Programmes. 2003. Available from: http://www.dh.gov.uk/assetRoot/04/02/30/60/04023060.pdf [Accessed March 3, 2007].
3 Lafata JE, Simpkins J, Lamerato L, et al. The economic impact of false-positive cancer screens. Cancer Epidemiol Biomarkers Prev 2004;13:2126–32.
4 Lidbrink E, Elfving J, Frisell J, Jonsson E. Neglected aspects of false positive findings of mammography in breast cancer screening: analysis of false positive cases from the Stockholm trial [see comments]. BMJ 1996;312:273–6.
5 Barton MB, Moore S, Polk S, et al. Increased patient concern after false-positive mammograms: clinician documentation and subsequent ambulatory visits. J Gen Intern Med 2001;16:150–6.
6 Brett J, Bankhead C, Henderson B, et al. The psychological impact of mammographic screening. A systematic review. Psychooncology 2005;14:917–38.
7 Brodersen J, Thorsen H, Cockburn J. The adequacy of measurement of short and long-term consequences of false-positive screening mammography. J Med Screen 2004;11:39–44.
8 Streiner DL, Norman GR. Health Measurement Scales: A Practical Guide to Their Development and Use. Oxford: Oxford University Press, 1995.
9 Doward LC, Meads DM, Thorsen H. Requirements for quality of life instruments in clinical research. Value Health 2004;7(Suppl.):S13–16.
10 McCaffery KJ, Barratt AL. Assessing psychosocial/quality of life outcomes in screening: how do we do it better? J Epidemiol Community Health 2004;58:968–70.
11 Cockburn J, De LT, Hurley S, Clover K. Development and validation of the PCQ: a questionnaire to measure the psychological consequences of screening mammography. Soc Sci Med 1992;34:1129–34.
12 Lightfoot N, Steggles S, Wilkinson D, et al. The short-term psychological impact of organised breast cancer screening. Curr Oncol (Toronto, 1198-0052) 1994;1:206–11.
13 Wright BD. Fundamental measurement for psychology. In: Embretson SE, Hershberger SL, eds. The New Rules of Measurement: What Every Psychologist and Educator Should Know. Mahwah, NJ and London: Lawrence Erlbaum Associates, 1999.
14 Andrich D. Controversy and the Rasch model: a characteristic of incompatible paradigms? Med Care 2004;42:I7–16.
15 Rosenbaum PR. Criterion-related construct validity. Psychometrika 1989;54:625–33.
16 Rasch G. An informal report on a theory of objectivity in comparisons. In: Van der Kamp LJTh, Vlek CAJ, eds. Psychological Measurement Theory. Leyden: University of Leyden, 1967.
17 Andersen EB. Sufficient statistics and latent trait models. Psychometrika 1977;42:69–81.
18 Bartholomew DJ. The Statistical Approach to Social Measurement. San Diego: Academic Press, 1996.
19 Holland PW, Hoskens M. Classical Test Theory as a first-order Item Response Theory: application to true-score prediction from a possibly nonparallel test. Psychometrika 2003;68:123–49.
20 Holland PW, Wainer H, (eds). Differential item functioning. Hilsdale NJ: Lawrence Erlbaum Associates, 1993.
21 Andrich D. Understanding resistance to the data-model relationship in Rasch's paradigm: a reflection for the next generation. J Appl Meas 2002;3:325–59.
22 Hunt SM, McEwen J, McKenna SP. Measuring health status: a new tool for clinicians and epidemiologists. J R Coll Gen Pract 1985;35:185–8.
23 Thorsen H, McKenna SP, Gottschalck L. The Danish version of the Nottingham Health Profile: its adaptation and reliability. Scand J Prim Health Care 1993;11:124–9.
24 Thorsen H, McKenna SP, Gottschalck L. Perceived health in three groups of elderly people. A validity study of the Danish version of the Nottingham Health Profile. Dan Med Bull 1995;42:105–8.
25 McDowell I, Newell C. Measuring Health. A Guide to Rating Scales and Questionnaires. Oxford: Oxford University Press, 1996.
26 Andrich D, Sheridan B, Luo G. RUMM2020. [Version 4.0 for Windows] 2005. RUMM Laboratory Pty Ltd. Available from: http://www.rummlab.com/ [Accessed March 5, 2007].

27 Andrich D, Luo G. Conditional pairwise estimation in the Rasch model for ordered response categories using principal components. J Appl Meas 2003;4:205–21.
28 Wright BD, Panchapakesan N. A Procedure for sample-free item analysis. Educ Psychol Meas 1969;29:23–48.
29 Cronbach LJ. Coeffcient alpha and the internal structure of tests. Psychometrika 1951;16:297–334.
30 Cronbach LJ. Internal consistency of tests: analyses old and new. Psychometrika 1988;53:63–70.
31 Smith EV Jr. Evidence for the reliability of measures and validity of measure interpretation: a Rasch measurement perspective. In: Smith EV Jr, Smith RM, eds. Introduction to Rasch Measurement. Theory, Models and Applications. Maple Grove: JAM Press, 2004.
32 Muthén L, Muthén B, Asparouhov T, Nguyen T. Mplus 2.14. 2005. Available from: http://www.statmodel.com/ [Accessed March 5, 2007].
33 Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Statist Soc Ser B 1995;57:289–300.
34 Masters GN. A Rasch model for partial credit scoring. Psychometrika 1982;47:149–74.
35 Andrich D. A rating formulation for ordered response categories. Appl Psychol Meas 1978;2:581–94.